Stereo Visual SLAM System in Underwater Environment

Sumei Pi, Bo He*, Shujing Zhang, Rui Nian,Yue Shen
Department of Electronics Engineering
Ocean University of China, Qingdao, China
pisumei520@163.com; bhe@ouc.edu.cn(*correspondence author); sjzhang365@gmail.com; nianrui_80@sina.com;
yue shen@ouc.edu.cn

Tianhong Yan
Department of Mechanical & Electrical Engineering
China Jiliang University
Hang Zhou, China
thyan@163.com

***Abstract*—With the increasing development of underwater vision sensors, simultaneous localization and mapping (SLAM) based on stereo vision has become a hot topic in the areas of ocean investigation and exploration. In this paper, visual SLAM with a focus on stereo camera system is presented to estimate the motion of autonomous underwater vehicles (AUVs) and build the feature map of surrounding environment in real-time. Feature detection and matching based on Speeded Up Robust Features (SURF) algorithm are implemented in the visual SLAM system. After eliminating the mismatch, we need to compute the stereo matched SURF features' local 3-D coordinates using the disparity values and stereo vision camera's parameters. Visual SLAM is implemented by fusing features coordinates and AUV pose with Extended Kalman Filter (EKF). The system has been verified on raw data gathered from the AUV in the underwater.**

Keywords—Visual SLAM; SURF; EKF; AUV;

I. INTRODUCTION

A robot computing a map of a previously unknown environment while localizing itself within that map is referred to as Simultaneous Localization and Mapping (SLAM). Recently, cameras have been used as the sensor yielding visual SLAM[1-4] which has increasingly become a hot topic for robot applications in recent times. When it comes to SLAM, cameras have become much more inexpensive than lasers, and also provide texture rich information about scene elements at any distance from the camera. Visual motion estimation techniques can provide very precise robot motion estimates. Finally, stable features can be detected in the images, fulfilling the data association functionality in a SLAM approach.

We herein present a stereo Visual SLAM system which can estimate the robot trajectory, meanwhile build and maintain a feature-based map using camera and other sensors. The camera traverses its environment yielding a trajectory. From each of the camera poses, a portion of the landmarks is observed. From these measurements, the most likely landmark positions and camera poses are estimated. Common methods designed to solve the estimation problem in real time rely on Extended Kalman Filters (EKFs). In the experiment, we employ EKF. And the map is usually represented as a set of features residing in 3D space. Therefore, feature detection and matching is the important process in the visual system. several papers consider the feature-based SLAM using camera images [5-7]. One herein must consider features that are as much invariant as possible with respect to any image transformation, when several unknown changes occur in the image, Point features are salient in images, have good invariant properties, and can be extracted with much less computation. A feature detection and matching algorithms include Scale Invariant Feature Transform (SIFT) [8], Speeded Up Robust Features (SURF) [9] and PCA-SIFT [10] etc.. In our case, one employ SURF algorithm to extract and match features. The SURF algorithm is presented in the paper, which is applied to an unknown environment based on binocular vision for SLAM; the simulation results are demonstrated to verify the feasibility and effectiveness of the algorithm.

The paper is organized as follows. In Section II, we present our system in more detail. Feature detection and matching algorithm and stereo image representation are introduced in Section III and Section IV respectively. Experimental results are given in section V.

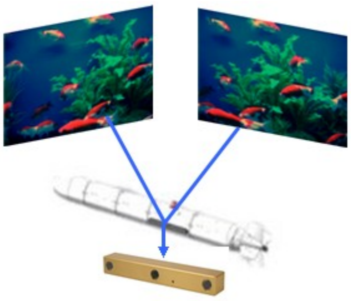

Fig. 1. The visual system

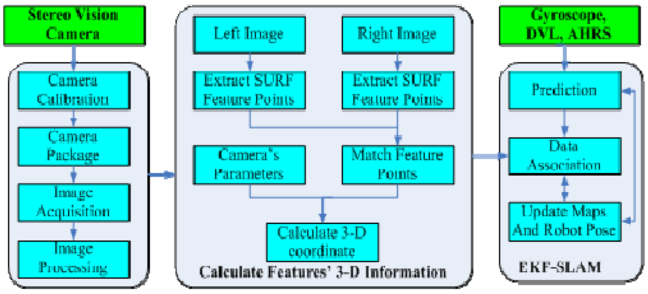

Fig. 2. The implementation process of vision SLAM

978-1-4799-3646-5/14/$31.00 ©2014 IEEE

II. STEREO VISUAL SLAM SYSTEM

In our simulation experiment, one kind of stereo vision camera, Bumblebee XB3 mounted in the AUV, was packaged to engage in the underwater image collection at regular intervals as well as the pre-calibrated against distortion and misalignment, as is shown in the Fig. 1. The underwater environment information based on camera and gyroscope, DVL(Global Position System) and AHRS (Attitude Heading Reference System) information can be fused by Extended Kalman Filter (EKF). The implementation process of visual SLAM algorithm is shown in Fig. 2. Data association in SLAM is the decision process of associating observations with existing features. When they are matched, the features may either (1) belong to a previously existing feature in the map or (2) be a new geometric feature or (3) be a spurious feature[11]. If the observations are existing features in the map, a correction of the state estimate will be done through the standard EKF update equations; if the observations are new features, then the system state vector will be augmented by adding the new observed features to the map to incrementally building the environment map; if the observations are determined as serious features, they will be rejected. To fulfill the data association functionality in a SLAM approach, stable features can be detected in the images. Feature detection and matching algorithm and stereo image representation are introduced in Section III and Section IV respectively. Features are obtained in every time step from the stereo images. Visual SLAM is implemented by fusing features coordinates and other sensors' information with Extended Kalman Filter (EKF).

III. SURF FEATURE MATCH ALGORITHM

In the experiment, we employ SURF algorithm to extract and match features. Implementation details are given further below.

A. SURF feature extraction

There are three steps to extract SURF feature: (1) Integral Images; (2) Fast-Hessian detector; (3) Interest Point Descriptor. Implementation details are given further below.

(a) The integral image is computed rapidly from an input image. Integral Image (summed area tables) is an intermediate representation for the image and contains the sum of gray scale pixel values of image. Using integral images is for major speed up.

(b) The SURF detector is based on the determinant of the Hessian matrix. Hessian determinant using the approximated Gaussians with box filters is referred to as the blob response at location $X = (x, y, \sigma)$. The simulation result of the responses is shown in Fig. 3. The search for local maxima of this function over both space and scale yields the interest points for an image. In order to detect interest points using the determinant of Hessian it is first necessary to introduce of a scale-space. A scale-space is used to find extrema across all possible scales. Fig. 4 illustrates the approach to constructing a scale-space in the SURF approach. It leaves the original image unchanged and varies only the filter size. In the experiment, the scale-space is divided into 5 octaves. Interest points can be extracted by thresholding, non-maximal suppression and interpolation.

c) Each interest point is assigned a reproducible orientation and a 64-dimensional vector. Fig. 5 (a) and (b) show the dominant orientation and description. The SURF descriptor image is showed in Fig. 6 (During the process of SURF feature detection, some important results of the left image from the stereo camera are showed)

B. Feature Match

Each frame in the binocular stereo vision can simultaneously get around two images of the same scene. It is presented in Fig. 7. Stereo vision matching is to find the 3-D points' corresponding relation of the same scene in two images. Features are matched between four images, namely the left and right images of two consecutive frames. Stable feature locations are achieved by matching features in a 'circle': The current left image is a reference image. Each feature from the current left image is independently matching against all features from the current right image, the previous left image and the previous right image by their Euclidean distance. SURF feature matching algorithm is done through the principle of the nearest neighbor proposed by Lowe [8]. The best candidate match for each feature is found by identifying the ratio of the closest neighbor and second-closest neighbor in the database of feature point (The closest neighbor is defined as the minimum Euclidean distance). When the ratio exceeds the threshold, the match is correct. Otherwise, it is incorrect. In the experiment, the threshold is 0.6. The result of initial matching is presented in Fig. 8. But there are incorrect matches in the matching progress.

To further improve the efficiency and accuracy of the matching, the mismatched feature points will be eliminated by the use of a homography. This can be described as follows:

$$\left| xH\hat{x} \right| \leq 1.5 \qquad (1)$$

Here x and $\hat{x}$ are the matching SURF feature points. Here H is the 3*3 homography matrix which maps x points in one image to $\hat{x}$ points in another. The final matching result is showed in Fig. 9. Fig. 8 and Fig. 9 show the result of any stereo image pairs.

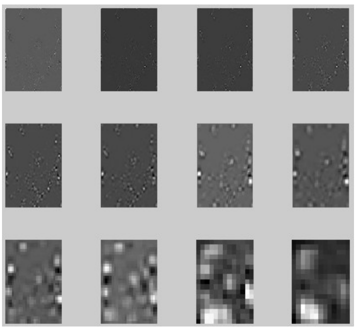

Fig. 3. Responses

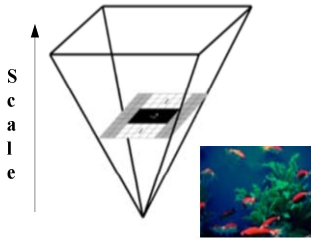

Fig. 4. box filters

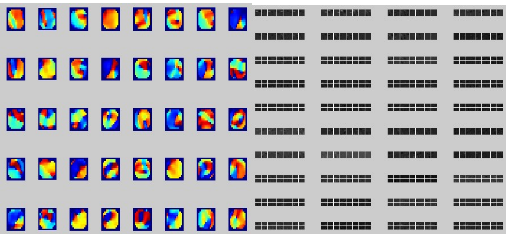

Fig. 5. (a) angle (dominant orientation) , (b) descriptor XY

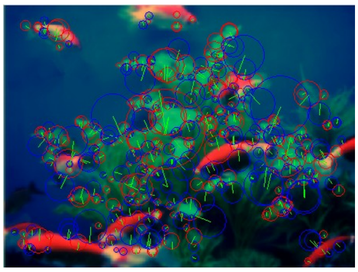

Fig. 6. SURF Descriptor

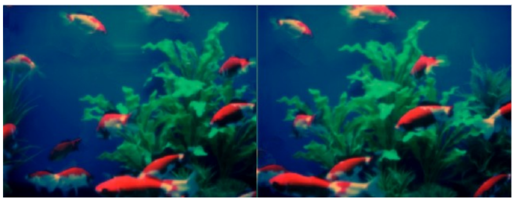

Fig. 7. Example underwater image pair

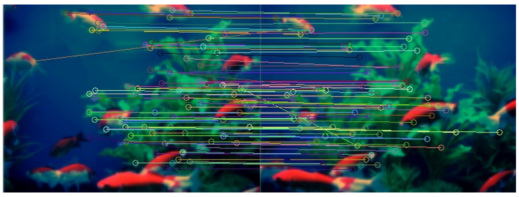

Fig. 8. Result of rough matching

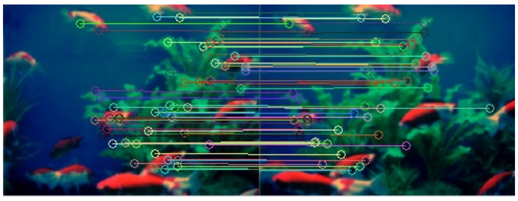

Fig. 9. Final matching result

IV. STEREO IMAGE REPRESENTATION

The section describes how to compute the features' coordinate. Implementation details are given further below.

There is one pair of underwater images I_R, I_L taken by our system. Let there be one spatial point P in the real world space, and the distance between the right projection centers O_R and the left projection centers O_L of the stereo camera is defined as the baseline distance B. The corresponding pixels of the left and right images taken from the spatial point P in the sea are represented as $P_L=(X_{left},Y_{right})$, $P_R=(X_{right},Y_{right})$. Fig.10 is the stereo imaging process in our system.

This section also describes the camera model used in the presented approach. The relationship between the camera frame C, with coordinate axes (x_c, y_c, z_c) and the homogeneous image coordinates I, with coordinate axes $(u,v,1)$ can be expressed as follows [12]:

$$(u,v,1)^{\mathrm{T}} = K \bullet (x_c, y_c, z_c)^{\mathrm{T}} \tag{2}$$

Let K be the 3 × 3 calibration matrix which encapsulates the intrinsic parameters of the camera. In particular, the 3-D points in the camera frame are given the disparity information provided by the stereo pair. Since the stereo camera provides rectified images, the backprojection equations to obtain a 3-D point are based on a pinhole camera model that relates image points and 3-D points using the following transformation function:

$$\begin{cases} x_c = \dfrac{B \bullet X_{left}}{Disparity} \\ y_c = \dfrac{B \bullet Y}{Disparity} \\ z_c = \dfrac{B \bullet f}{Disparity} \end{cases} \tag{3}$$

Where B is the baseline distance, and f is the focal length. Because the cameras are horizontally aligned, the distances from the top of the image to the matching features are exactly the same in both images, $Y = Y_{left} = Y_{right}$. And the disparity of point P is $Disparity = X_{left} - X_{right}$. Equation (3) is the concrete application of equation (2). In general, the camera frame and the world frame are not aligned, but the two coordinate frames are related via a translation vector t and a rotation matrix R, the extrinsic calibration of the camera. Given a 3-D point $X_W = (x_w, y_w, z_w)^{\mathrm{T}}$ in the world reference frame, the corresponding point $X_C = (x_c, y_c, z_c)^{\mathrm{T}}$ in the camera coordinate frame is computed via:

$$X_C = R \bullet X_W + t \tag{4}$$

Combining equations (2) and (4) the mapping of a 3-D object point onto the image plane is described as:

$$(u,v,1)^{\mathrm{T}} = P \bullet (x_w, y_w, z_w, 1)^{\mathrm{T}} \tag{5}$$

Here $P = K \bullet [R|t]$ is a 3 × 4 projection matrix [12].

In our experiment, 3-D points are computed using the equations (2), (3), (4) and (5). These points are the nature landmarks. Local 3-D points in the robot coordinate system from a frame image is shown in Fig. 11.

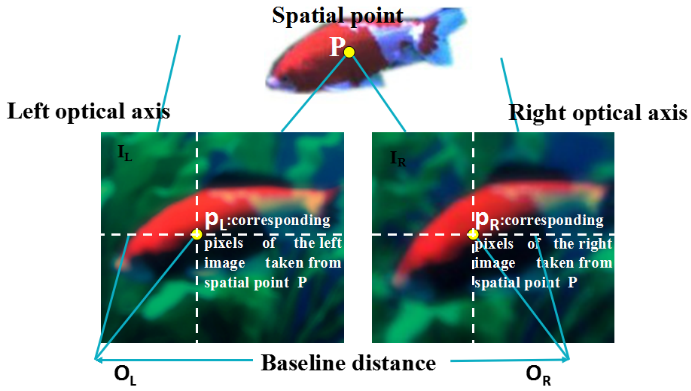
Spatial point
P
Left optical axis
Right optical axis
I_L
I_R
p_L:corresponding pixels of the left image taken from spatial point P
p_R:corresponding pixels of the right image taken from spatial point P
O_L
Baseline distance
O_R

Fig. 10. Stereo imaging process

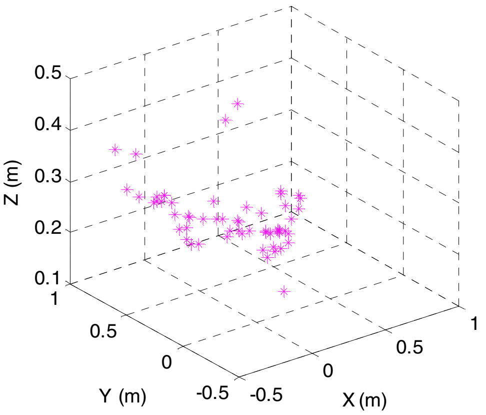

Fig. 11. Local 3-D coordinates

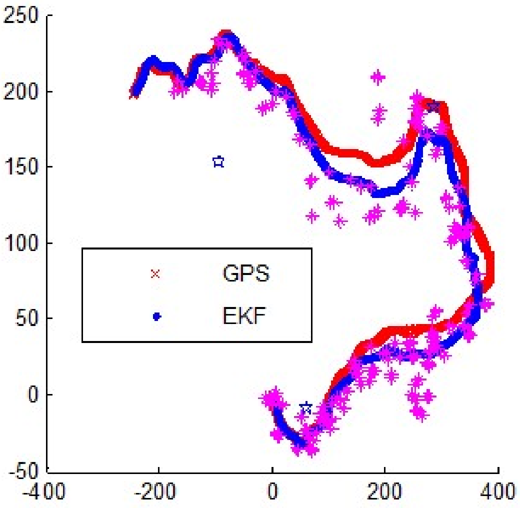

Fig. 12. EKF-SLAM

V. Visual SLAM Result

SLAM algorithm can estimate the robot trajectory, meanwhile build and maintain a feature-based map. The SLAM algorithm we use is based on EKF (Extended Kalman Filter). Visual SLAM is completed by fusing the information of stereo vision and robot pose with EKF. EKF is usually used to estimate an augmented state constituted by the vehicle pose and landmark positions. The vehicle travels through the environment using its sensors to generate a prediction by the vehicle motion model and using the measurement sensors mounted on the vehicle to observe features around it. System equation and measurement equation are expressed as follows:

$$X_{k+1} = f(X_k) + q_k \tag{6}$$

$$Z_{k+1} = h(X_{k+1}) + r_{k+1} \tag{7}$$

where f (.) is the non-linear system equation, h (.) is the non-linear measurement equation described above. X_k is the state of the system at time step k, and the state of the system is represented by an augmented vector including the state of the vehicle and the state of a set of N features of the environment, $X_k = [X_v, X_m]^T$. In our case, the state of the vehicle X_v and the vector of the map X_m can be described as:

$$X_v = [x_v, y_v, z_v, \Phi_v, \varphi_v, \theta_v]^T \tag{8}$$

$$X_m = [x_1, y_1, z_1, \ldots, x_N, y_N, z_N]^T \tag{9}$$

$Z_{k+1} = [u_{R,k+1,1}, v_{R,k+1,1}, \ldots, u_{R,k+1,N}, v_{R,k+1,N}]^T$ denotes the 4N dimensional measurement vector, where N denotes the number of feature correspondences used for filtering. q_k and r_{k+1} are the system noise and the measurement noise respectively, which are assumed to be uncorrelated.

In the experiments, we simulate the process of visual SLAM algorithm using real data sets and equation (6)-(7) by subsequent iteration. The final partial result of underwater experiment is showed in Fig. 12. Supposing the vehicle no floating up and down, the result is presented in the XOZ plane where the red and blue lines represent the GPS data and SLAM trajectory respectively, and the purple asterisks represent the environmental landmarks.

Conclusion

In this paper, visual SLAM with a focus on stereo camera system is presented to estimate the motion of autonomous underwater vehicles (AUVs) and the map of surrounding environment in real-time. The system has been implemented on raw data gathered from a AUV. It has been shown that in the simulation experiments the presented approach could get good performance in effectiveness. Herein we present a sparse SLAM, these sparse map representations are often insufficient for tasks common such as path planning, collision avoidance. In the future we plan to research a dense stereo visual SLAM system in the underwater.

ACKNOWLEDGMENT

This work is partially supported by the High Technology Research and Development Program of China (2006AA09Z231), Natural Science Foundation of China (41176076, 31202036, 51075377, 51379198), the Science and Technology Development Program of Shandong Province (2008GG1055011, BS2009HZ006), and the Key Science & Technology Project of Shandong Province(2013GHY11507).

REFERENCES

[1] Mei C, Sibley G, Cummins M, et al. RSLAM: A system for large-scale mapping in constant-time using stereo[J]. International journal of computer vision, 2011, 94(2): 198-214.

[2] Thomas S J. Real-time stereo visual slam[J]. Master's thesis, Heriot-Watt University, Universitat de Girona, Universite de Bourgogne, 2008.

[3] Artieda J, Sebastian J M, Campoy P, et al. Visual 3-d slam from uavs[J]. Journal of Intelligent and Robotic Systems, 2009, 55(4-5): 299-321.

[4] Muhammad N, Fofi D, Ainouz S. Current state of the art of vision based SLAM[C]//IS&T/SPIE Electronic Imaging. International Society for Optics and Photonics, 2009: 72510F-72510F-12.

[5] Wang P L, Shi S D, Hong X W. A SLAM algorithm based on monocular vision and odometer[J]. Comput Simul, 2008, 25: 172-175.

[6] Davison A J. Real-time simultaneous localisation and mapping with a single camera[C]//Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on. IEEE, 2003: 1403-1410.

[7] Kim G H, Kim J S, Hong K S. Vision-based simultaneous localization and mapping with two cameras[C]//Intelligent Robots and Systems, 2005.(IROS 2005). 2005 IEEE/RSJ International Conference on. IEEE, 2005: 1671-1676.

[8] Lowe D G. Distinctive image features from scale-invariant keypoints[J]. International journal of computer vision, 2004, 60(2): 91-110.

[9] Bay H, Tuytelaars T, Van Gool L. Surf: Speeded up robust features[M]//Computer Vision–ECCV 2006. Springer Berlin Heidelberg, 2006: 404-417.

[10] Ke Y, Sukthankar R. PCA-SIFT: A more distinctive representation for local image descriptors[C]//Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on. IEEE, 2004, 2: II-506-II-513 Vol. 2.

[11] Zhang S, Xie L, Adams M. An efficient data association approach to simultaneous localization and map building[J]. The International Journal of Robotics Research, 2005, 24(1): 49-60.

[12] Hartley R, Zisserman A. Multiple view geometry in computer vision[M]. Cambridge university press, 2003.